Convex Optimisation Homework Solutions

---

1. (a) Suppose first that x and y are both non-negative. Then $f(x) = x$, $f(y) = y$, and it is straightforward that, for all $\alpha \in [0, 1]$,

$$f(\alpha x + (1-\alpha)y) = \alpha x + (1-\alpha)y = \alpha f(x) + (1-\alpha)f(y).$$

Thus, the condition for convexity is satisfied in this case. The case when x and y are both negative is also easy. The only case remaining to consider is when one is positive and the other negative. Suppose that $x < 0$ and $y > 0$, and let $\alpha in[0, 1]$. Now,

$$\begin{aligned} f(\alpha x + (1-\alpha)y) = |\alpha x + (1-\alpha)y| &\leq |\alpha x| + |(1-\alpha)y| \\ &= \alpha|x| + (1-\alpha)|y| = \alpha f(x) + (1-\alpha)f(y). \end{aligned}$$

The inequality in the above equation is called the triangle inequality, and is known to hold for all real (and complex) numbers.

(b) The function f is defined on a convex domain, the open interval $(0, \infty)$, and is twice differentiable on this domain. So we can check convexity using the criterion that the second derivative should be non-negative everywhere. We have

$$f'(x) = 1 + \log x, \quad f''(x) = \frac{1}{x} > 0 \; \forall \; x > 0.$$

(c) We again use the second derivative criterion, now in matrix form. The Hessian matrix of f is

$$D^2 f(x, y) = 2 \begin{pmatrix} y^2 & 2xy \\ 2xy & x^2 \end{pmatrix},$$

and so, we obtain for any $a, b \in \mathbb{R}$ that

$$\begin{pmatrix} a & b \end{pmatrix} D^2 f(x, y) \begin{pmatrix} a \\ b \end{pmatrix} = 2(a^2y^2 + 2(a+b)xy + b^2x^2) = 2(ay + bx)^2 \geq 0.$$

Hence, $D^2 f(x, y)$ is positivie semi-definite for any (x, y), so f is convex.

(d) The approach is the same as for the last part, to compute the Hessian of f. Differentiating once, we get

$$\nabla f(\mathbf{x}) = 2(A\mathbf{x} - \mathbf{b})^T A.$$

To check this, observe that

$$\begin{aligned} & f(\mathbf{x} + \epsilon\mathbf{y}) - f(\mathbf{x}) - \epsilon\nabla f(\mathbf{x})\mathbf{y} \\ &= (A\mathbf{x} + \epsilon A\mathbf{y} - \mathbf{b})^T(A\mathbf{x} + \epsilon A\mathbf{y} - \mathbf{b}) - (A\mathbf{x} - \mathbf{b})^T(A\mathbf{x} - \mathbf{b}) - 2\epsilon(A\mathbf{x} - \mathbf{b})^T A\mathbf{y} \\ &= \epsilon^2\mathbf{y}^T A^T A\mathbf{y} = o(\epsilon). \end{aligned}$$

To obtain the second equality, we have used the fact that $\mathbf{y}^T A^T(A\mathbf{x} - \mathbf{b})$ is a scalar, and hence equal to its transpose, which is $(A\mathbf{x} - \mathbf{b})^T A\mathbf{y}$. Differentiating once more,

$$D^f(\mathbf{x}) = A^T A \quad \forall \; \mathbf{x} \in \mathbb{R}^n.$$

But this matrix is positive semidefinite because, for all $\mathbf{z} \in \mathbb{R}^n$, we have

$$\mathbf{z}^T A^T A\mathbf{z} = \|A\mathbf{z}\|^2 \geq 0.$$

Hence, f is convex.

(e) If x_0 and x_1 are both strictly positive, or one is positive and the other is zero, then it is easy to check the condition that

$$f((1-\alpha)x_0 + \alpha x_1) \leq (1-\alpha)f(x_0) + \alpha x_1;$$

the condition holds with equality if both are strictly positive, and with strict inequality if one is positive and the zero. If one of the them is positive and the other negative, the right hand side of the inequality in the equation above is $+\infty$, so the inequality holds irrespective of the value of the left hand side.

2. (a) Suppose $A, B \in \mathbb{S}$, i.e., A and B are both symmetric matrices. Then $a_{ij} = a_{ji}$ and $b_{ij} = b_{ji}$ for all $i, j \in \{1, 2, \ldots, n\}$. Now, fix $\alpha in[0,1]$ and let $C = \alpha A + (1-\alpha)B$. Then,

$$c_{ij} = \alpha a_{ij} + (1-\alpha)b_{ij} = \alpha a_{ji} + (1-\alpha)b_{ji} = c_{ji}$$

for all $i, j \in \{1, 2, \ldots, n\}$, i.e., C is symmetric. Thus, we have shown that $\mathbb{S}$ is convex.

(b) By part (a), the function f is defined on a convex domain. We now use the hint. Observe that, for any fixed $\mathbf{x} \in \mathbb{R}^n$, the map $A \mapsto \mathbf{x}^T A\mathbf{x}$ is linear (i.e, if we define $g(A) = \mathbf{x}^T A\mathbf{x}$, then $g(A+B) = g(A) + g(B)$ for any two symmetric matrices A and B, and $g(cA) = cg(A)$ for any symmetric matrix A, and any real number c). But linear functions are convex, so $A \mapsto \mathbf{x}^T A\mathbf{x}$ is convex. Now, by the Rayleigh-Ritz formula, $\lambda_{\max}(A)$ is simply the maximum of these convex functions indexed by $\mathbf{x}$ over $\{\mathbf{x} : \|\mathbf{x}\| = 1\}$, and we know that the maximum of convex functions is convex.

3. Let x_0, x_1 and x_α be as in the hint, and so also for y_0, y_1 and y^α. By definition of these, we have

$$g(x_0) = f(x_0, y_0), \quad g(x_1) = f(x_1, y_1), \quad g(x_\alpha) = f(x_\alpha, y^\alpha) \leq f(x_{\alpha)}, y_\alpha). \tag{1}$$

We also have by the convexity of f that

$$f(x_\alpha, y_\alpha) \leq \alpha f(x_0, y_0) + (1-\alpha)f(x_1, y_1).$$

Substituting the above inequality in equation (1), we obtain $g(x_\alpha) \leq \alpha g(x_0) + (1-\alpha)g(x_1)$. Thus, we have shown that g is convex.

4. (a) We have already shown that f is convex in problem 1(d); only the notation has changed. The first order condition for minimality of a convex function is that its gradient is zero. So we want

$$\nabla f(\beta) = 2X^T(X\beta - \mathbf{y}) = \mathbf{0}, \quad \text{i.e.,} \quad \beta = \left(X^T X\right)^{-1} X^T \mathbf{y}.$$

(b) We showed in 1(a) that $|x|$ is a convex function of $x \in \mathbb{R}$. Much the same argument shows that for any fixed i, $|\beta_i|$ is a convex function of $\beta \in \mathbb{R}^n$; the co-ordinates β_j, $j \neq i$, play no role. Now, $g(\beta)$ is a linear combination of the convex functions $\|(X\beta - \mathbf{y}\|^2$ and $|\beta_i|$, $i = 1, 2, \ldots, n$ with positive coefficients $1, \lambda, \lambda, \ldots, \lambda$. Hence, it is convex.

5. (a) The objective function is $f_0(\mathbf{x}) = \frac{1}{2}\mathbf{x}^T Q\mathbf{x} + \mathbf{c}^T\mathbf{x}$, and its second derivative matrix is $D^2 f_0(\mathbf{x}) = Q$. As we are given that Q is p.s.d., it follows that f_0 is convex.

Next, suppose that $\mathbf{x}$ and $\mathbf{y}$ in $\mathbb{R}^n$ both satisfy the constraints. Let $\alpha \in [0,1]$. Then,

$$A(\alpha\mathbf{x} + (1-\alpha)\mathbf{y}) = \alpha A\mathbf{x} + (1-\alpha)A\mathbf{y} \le \alpha\mathbf{b} + (1-\alpha)\mathbf{b} = \mathbf{b},$$

i.e., $\alpha\mathbf{x}+(1-\alpha)\mathbf{y}$ also satisfies the constraints. Thus, the feasible set is convex, and we are minimising a convex function over a convex set. This is the definition of a convex optimisation problem.

(b) There are no equality constraints in this problem. Thus, the Lagragian is

$$L(\mathbf{x},\lambda) = \frac{1}{2}\mathbf{x}^T Q\mathbf{x} + \mathbf{c}^T\mathbf{x} + \lambda^T(A\mathbf{x} - \mathbf{b}), \quad \mathbf{x} \in \mathbb{R}^n, \lambda \in \mathbb{R}^m_+.$$

The dual objective function is given by $g(\lambda) = \min_{\mathbf{x}\in\mathbb{R}^n} L(\mathbf{x},\lambda)$. As L is convex in $\mathbf{x}$ for any fixed λ, we can use the first order sufficient conditions for optimality, which are:

$$\nabla_{\mathbf{x}} L(\mathbf{x},\lambda) = \mathbf{x}^T Q + \mathbf{c}^T + \lambda^T A = \mathbf{0}.$$

We will assume that Q is invertible, in which case the above equation has solution $\mathbf{x} = -Q^{-1}(\mathbf{c} + A^T\lambda)$. Substituting this in the expression for g, and using the fact that Q is symmetric, i.e. $Q^T = Q$, we obtain that

$$\begin{aligned} g(\lambda) &= \frac{1}{2}(\mathbf{c} + A^T\lambda)^T Q^{-1}(\mathbf{c} + A^T\lambda) - (\mathbf{c} + A^T\lambda)^T Q^{-1}(\mathbf{c} + A^T\lambda) + \lambda^T\mathbf{b} \\ &= -\frac{1}{2}(\mathbf{c} + A^T\lambda)^T Q^{-1}(\mathbf{c} + A^T\lambda) + \lambda^T\mathbf{b}. \end{aligned}$$

Thus, the dual problem is:

$$\max_{\lambda\in\mathbb{R}^m_+} -\frac{1}{2}(\mathbf{c} + A^T\lambda)^T Q^{-1}(\mathbf{c} + A^T\lambda) + \lambda^T\mathbf{b},$$

or equivalently, ignoring the constant term $\mathbf{c}^T Q^{-1}\mathbf{c}$,

$$\min_{\lambda\in\mathbb{R}^m} \frac{1}{2}\lambda^T\big(AQ^{-1}A^T\big)\lambda + \big(AQ^{-1}\mathbf{c} - \mathbf{b}\big)^T\lambda, \quad \text{subject to } \lambda \ge \mathbf{0},$$

which we recognise as another QP. Moreover, if Q is positive definite, so is Q^{-1} and $AQ^{-1}A^T$ is p.s.d. (exercise), so the dual problem is also a convex program.

(c) The KKT conditions for optimality are:

$$\mathbf{x}^T Q + \mathbf{c}^T + \lambda^T A = \mathbf{0}^T, \quad \lambda^T(A\mathbf{x} - \mathbf{b}) = 0,$$

where the latter are the complementary slackness conditions.

(d) The gradient of the objective function at $\mathbf{x}^0$ is

$$\nabla f_0(\mathbf{x}^0) = (\mathbf{x}^0)^T Q + \mathbf{c}^T.$$

So, for gradient descent, we are seeking $\mathbf{x}^1$ of the form

$$\mathbf{x}^1 = \mathbf{x}^0 - t(Q\mathbf{x}^0 + \mathbf{c}).$$

With exact line search, we want the value of t that exactly minimises

$$f_0(\mathbf{x}^1) = f_0(\mathbf{x}^0) - t(\mathbf{x}^0)^T Q(Q\mathbf{x}^0 + \mathbf{c}) + \frac{t^2}{2}(Q\mathbf{x}^0 + \mathbf{c})^T Q(Q\mathbf{x}^0 + \mathbf{c}) - t\mathbf{c}^T(Q\mathbf{x}^0 + \mathbf{c}).$$

Ignoring the term $f_0(\mathbf{x}^0)$ that does not depend on t, and using the fact that Q is symmetric, we see that the objective can be rewritten as minimising the following quadratic function of t:

$$h(t) = \frac{t^2}{2}(Q\mathbf{x}^0 + \mathbf{c})^T Q(Q\mathbf{x}^0 + \mathbf{c}) - t(Q\mathbf{x}^0 + \mathbf{c})^T (Q\mathbf{x}^0 + \mathbf{c}).$$

Now, Q is positive definite, so the coefficient of t^2 is non-negative. Hence, the minimum is attained at

$$t^* = \frac{(Q\mathbf{x}^0 + \mathbf{c})^T (Q\mathbf{x}^0 + \mathbf{c})}{(Q\mathbf{x}^0 + \mathbf{c})^T Q(Q\mathbf{x}^0 + \mathbf{c})}.$$

If the denominator is zero, the function $h(t)$ is linear, the minimum value of $h(t)$ is $-\infty$. But, as Q is positive definite, this can only happen if $Q\mathbf{x}^0 + \mathbf{c} = \mathbf{0}$, i.e., $\mathbf{x}^0 = Q^{-1}\mathbf{c}$, which is the solution of the QP. Hence, either $\mathbf{x}^0$ is already the solution, or gradient descent with exact line search takes us to the new value

$$\mathbf{x}^1 = \mathbf{x}^0 - t^*(Q\mathbf{x}^0 + \mathbf{c}),$$

where t^* is as above. The update step involves 3 multiplications of an n-vector by an $n \times n$ matrix, as well as some vector additions, which are comparatively cheap. The matrix multiplication involves n^2 scalar multiplications and dominates the cost. Thus, the computational complexity of each iteration is $O(n^2)$.

By comparsion, a step of the Newton algorithm takes us to

$$\tilde{\mathbf{x}}^1 = \mathbf{x}^0 - \left(D^2 f(\mathbf{x}^0)\right)^{-1} \nabla f(\mathbf{x}^0)^T.$$

Noting that $D^2 f(\mathbf{x}^0) = Q$, and using the value of $\nabla f(\mathbf{x}^0)$ computed above, we get

$$\tilde{\mathbf{x}}^1 = \mathbf{x}^0 - Q^{-1}(Q\mathbf{x}^0 + \mathbf{c}) = -Q^{-1}\mathbf{c}.$$

But the latter is exactly where the minimum is attained, as can be seen by solving $\nabla f(\mathbf{x}) = \mathbf{0}$. Hence, for unconstrained quadratic programs, a single step of the Newton method takes us to the solution, irrespective of the starting condition. This is the intuition motivating the Newton method, that, in the vicinity of a minimum, a twice-differentiable convex function looks approximately like a quadratic function.

In terms of computational complexity, solving n equations in n unknowns by any of the standard algorithms, such as LU or QR decomposition, requires $O(n^3)$ computations in the worst case. This compares with $O(n^2)$ for an iteration of gradient descent. So there are advantages and disadvantages for both methods.